ISSN: 2664-9640

Exporing the Potential of Big Data Technology for Early Warning System on Academic Performance of Students

Zihan Wu, Ruping Zhang

School of Management Science and Engineering, Anhui University of Finance and Economics, Anhui 233000, China;

Abstract

The current trend towards increased levels of informatization in modern society, particularly in conjunction with the utilization of big data and data visualization technologies, has had a profound impact on the field of education. As a result, most educational institutions have developed academic systems that are capable of storing student test data. Despite this progress, the majority of schools' performance evaluation systems remain limited to querying basic metrics such as grades and average scores. Consequently, a substantial amount of potentially valuable information is being lost amidst the vast quantities of available data. This project represents an innovative response to the ongoing educational revolution, seeking to explore the feasibility of leveraging big data technologies to provide more nuanced student grade warnings. Through a review of recent literature and the implementation of novel student grade warning code, this project offers a series of constructive recommendations for future improvements in this area.

Keywords

Python, Data Analysis, Performance Alerting.

1. Background

In the field of analyzing and predicting student performance, researchers around the world have utilized big data techniques to study and analyze data related to learning behavior in order to construct prediction models. For instance, Marbouti et al.[1] employed Bayesian classifiers and integrated models to predict student performance and identify those who are at academic risk. Oyerinde[2] used linear regression models to predict students' academic performance. Lopez[3] used data from teaching platforms as predictive factors for predicting student grades. In China, Junmin Ye et al. proposed a short-text-based emotionally enhanced performance prediction method, which predicts learners' performance based on their learning status. Song Dan et al.[4] collected and analyzed multiple sources of data generated during the teaching process to explore students' learning status and effectiveness and predict their course learning status for early warning[5]. These studies highlight the importance of using big data techniques and predictive models to better understand student performance and provide insights that can inform educational practices and policies.

2. K-Nearest Neighbors and ROC curve

The k-nearest neighbor (KNN) algorithm is a widely used classification technique due to its simplicity, effectiveness, and suitability for automatic classification of large datasets. Its low complexity also makes it an attractive option compared to other algorithms in the field of data mining. Since its introduction in 1967, KNN has gained significant popularity as a research tool, serving as a reliable and efficient classification method for various applications.

ISSN: 2664-9640

The KNN algorithm can be succinctly summarized by the following steps: (i) selecting the number of neighbors, K, and the appropriate distance metric, (ii) identifying the K nearest neighbors in the sample with the relevant classification, and (iii) utilizing a majority voting scheme based on the class labels of the nearest neighbors to make the final classification decision. These steps effectively highlight the fundamental components of the KNN algorithm and demonstrate its practical utility as a classification technique in the field of data mining. As shown in Figure 1.

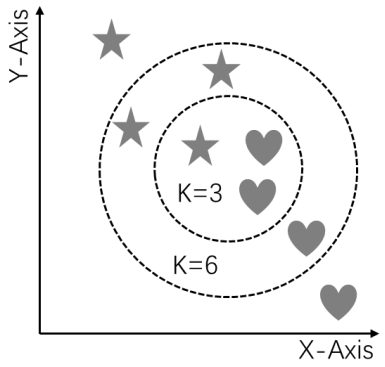

Figure 1: Example of KNN Algorithm

For the problem of predicting student performance, we can treat it as a binary classification problem, where students with grades greater than or equal to 60 are labeled as positive samples, and students with grades less than 60 are labeled as negative samples. In order to evaluate the performance of the prediction model, ROC curves can be used, which plot the false positive rate against the true positive rate to determine the accuracy of the model. In student performance prediction, the optimal threshold value represents the best score line for predicting performance. For example, if the optimal threshold is 70, students whose predicted grades are greater than or equal to 70 are labeled as positive samples, and those whose predicted grades are less than 70 are labeled as negative samples. By adjusting the threshold, we can balance the accuracy and recall of the model and choose the most appropriate threshold. This approach can improve the accuracy and reliability of the prediction and be beneficial for academic research and practical applications.

3. Experiment

The aim of this study is to effectively preprocess student achievement data using big data technology and establish a high-quality student achievement database. Based on this, we employ machine learning techniques to construct a predictive model for student achievement trends in the absence of intervention. Additionally, by introducing the receiver operating characteristic (ROC) curve method, we can identify students with abnormal performance based on a predetermined threshold and provide timely warnings. Our goal is to analyze the feasibility of big data technology in student achievement warning systems and propose beneficial suggestions through this research.

3.1. Data processing

Firstly, we utilize the pandas library to import the student grade data for each subject into a pandas Dataframe. We then preprocess the data by removing any rows that contain non-integer grades to ensure the data's accuracy and integrity. Finally, we save the cleaned data and utilize a Python script to batch process the grades of six subjects in three exams. This comprehensive data cleaning and processing methodology ensures that the data is reliable and suitable for subsequent analysis and modeling, providing a solid foundation for academic research and practical applications.

ISSN: 2664-9640

3.2. Predictive Models and Warning

We proceeded to develop a K-nearest neighbor (KNN) regression model for predicting student achievement. The code initially utilized the Pandas tool to read in the pre-processed achievement data and subsequently partitioned it into training and test sets. Thereafter, data normalization was applied to the feature matrix to enhance the model's precision. Subsequently, a prediction model was established using the KNN regression algorithm and tested on the test set. The prediction results and true values were then printed out and visualized to showcase the predictive performance of the model. Figure 2 displays selected results.

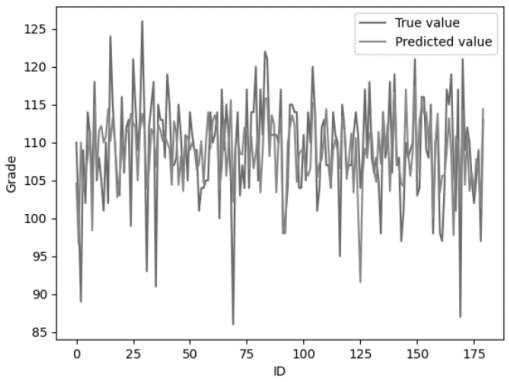

Figure 2: Partial Prediction Results of Language Performance for Senior

Once the model is built, we can use it to make predictions on a test set of data. We then plot the ROC curve using the true positive rate (TPR) and false positive rate (FPR) values calculated from the model's predictions.

The ROC curve helps us evaluate the model's performance by showing how well it can distinguish between positive (at-risk) and negative (not at-risk) cases. The closer the curve is to the upper left corner of the graph, the better the model's performance.

To determine the threshold for the warning score, we look for the point on the ROC curve where the TPR and FPR are balanced. This is known as the "knee" of the curve. The threshold score is then set at the point on the curve closest to the knee.

Once we have determined the threshold, we can use it to predict which students are at risk of poor performance and provide them with targeted interventions to improve their academic outcomes.

3.3. Areas for improvement

(1) Insufficient data quantity leads to inaccurate prediction results as it fails to adequately cover the long-term performance trends of high school students.

(2) Poor data quality, including noise or outliers, undermines the stability and reliability of prediction results.

(3) The selection of an inappropriate K value may hinder the ability to achieve optimal prediction performance and requires optimization and adjustment.

(4) Inappropriate methods or parameters selected for data normalization can compromise the accuracy and stability of the model, necessitating optimization and adjustment to improve its performance.

4. Conclusion

In summary, this project demonstrates the potential and feasibility of using big data technology to improve student performance prediction systems. By collecting and analyzing student performance data and utilizing machine learning methods to study performance trends and attempt to define threshold values for early warning systems. Building predictive analytic

ISSN: 2664-9640

models is another application of big data technology, which has been extensively researched in recent years. As more data becomes available and machine learning algorithms improve, we can develop more accurate and sophisticated early warning models in the future. This will enable educators to better understand student needs and identify students with declining performance trends, allowing for targeted interventions to improve student performance.

Acknowledgements

This project is sponsored by the Research Fund of Anhui University of Finance and Economics. (No.XSKY22146, Research on Early Warning System of High School Students' Performance Based on Big Data Technology)

References

[1] Marbouti F, Diefes-Dux H A, Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading[J]. Computers & Education, 2016, 103: 1-15.

[2] Oyerinde O D, Chia P A. Predicting students' academic performances–A learning analytics approach using multiple linear regression[J]. 2017.

[3] López S L S, Redondo R D, Vilas A F. Predicting students' grade based on social and content interactions[J]. The International journal of engineering education, 2018, 34(3): 940-952.

[4] Junmin Ye, Daxiong Luo,Chen Shu. A short text-based sentiment enhancement method for predicting online learners' performance[J]. Journal of Automation, 2020, 46(9): 1927-1940.

[5] Song Dan, Liu Dongbo, Feng Xia. Research on course grade prediction and course early warning based on multi-source data analysis[J]. Research on Higher Engineering Education, 2020, 1.

[6] Sun J. Research on abnormal performance detection based on FAST-MCD algorithm[J]. Modern Computer,2021,27(29):59-62.

[7] Xu Shengdong. Analytical study of student achievement prediction model based on big data technology[J]. Popular Standardization,2021(12):51-53.

[8] Yang S. H.,Li M.. A method for detecting abnormal scores based on distribution features[J]. Journal of South China University (Natural Science Edition),2008,22(04):7-9+21.

[9] Li M., Yang S. H.. A distance-based method for detecting abnormal scores[J]. Journal of South China University (Natural Science Edition),2009,23(04):70-73+78.

[10] Zhang Bingzhu, Li Hao, Hou Hexiang, Yu Haitao, Ma Xingguang. Design of early warning system for student performance in universities based on database and machine learning technology[J]. Chinese medicine education,2021,40(03):63-67.

[11] Lin, Mengnan, Li, Jinhui. A neural network model for student grade prediction based on adaptive differential evolution[J]. Modern Electronics Technology,2022,45(03):130-134.

[12] Ma Hao-Xuan,Yang Xiao-Qian,Liao Zhen,Cui Chen,Tension Dan. Research on temporal data anomaly detection algorithm[J]. Science and Technology Innovation,2022(06):78-81.

[13] Gao Hongmei,Wei Long. Research and reflection on the application of data mining in university academic affairs management[J]. Technology and Market,2021,28(04):131-134+137.